Coding with Constraints: Minimum Distance Bounds and Systematic Constructions

Wael Halbawi, Matthew Thill & Babak Hassibi
Department of Electrical Engineering
California Institute of Technology
Pasadena, California 91125
Email: {whalbawi,mthill,hassibi}@caltech.edu

Abstract

We examine an error-correcting coding framework in which each coded symbol is constrained to be a function of a fixed subset of the message symbols. With an eye toward distributed storage applications, we seek to design systematic codes with good minimum distance that can be decoded efficiently. On this note, we provide theoretical bounds on the minimum distance of such a code based on the coded symbol constraints. We refine these bounds in the case where we demand a systematic linear code. Finally, we provide conditions under which each of these bounds can be achieved by choosing our code to be a subcode of a Reed-Solomon code, allowing for efficient decoding. This problem has been considered in multisource multicast network error correction. The problem setup is also reminiscent of locally repairable codes.

I. Introduction

We consider a scenario in which we must encode s message symbols using a length n error-correcting code subject to a set of encoding constraints. Specifically, each coded symbol is a function of only a subset of the message symbols. This setup arises in various situations such as in the case of a sensor network in which each sensor can measure a certain subset of a set of parameters. The sensors would like to collectively encode the readings to allow for the possibility of measurement errors. Another scenario is one in which a client wishes to download data files from a set of servers, each of which stores information about a subset of the data files. The user should be able to recover all of the data even in the case when some of the file servers fail. Ideally, the user should also be able to download the files faster in the absence of server failures. To protect against errors, we would like the coded symbols to form an error-correcting code with reasonably high minimum distance. On the other hand, efficient download of data is permitted when the error-correcting code is of systematic form. Therefore, in this paper, we present an upper bound on the minimum distance of an error-correcting code when subjected to encoding constraints, reminiscent of the cut-set bounds presented in [1]. In certain cases, we provide a code construction that achieves this bound. Furthermore, we refine our bound in the case that we demand a systematic linear error-correcting code, and present a construction that achieves the bound. In both cases, the codes can be decoded efficiently due to the fact that our construction utilizes Reed-Solomon codes.

A. Prior Work

The problem of constructing error-correcting codes with constrained encoding has been addressed by a variety of authors. Dau et al. [2], [3], [4] considered the problem of finding linear MDS codes with constrained generator matrices. They have shown that, under certain assumptions, such codes exist over large enough finite fields, as well as over small fields in a special case. A similar problem known as the weakly secure data exchange problem was studied in [5],[6]. The problem deals with a set of users, each with a subset of messages, who are interested in broadcasting their information securely when an eavesdropper is present. In particular, the authors of [6] conjecture the existence of secure codes based on Reed-Solomon codes and present a randomized algorithm to produce them. The problem was also considered in the context of multisource multicast network coding in [1], [7], [8]. In [7], the capacity region of a simple multiple access network with three sources is achieved using Reed-Solomon codes. An analogous result is derived in [8] for general multicast networks with 3 sources using Gabidulin codes.

There has been a recent line of work involving codes with local repairability properties, in which every parity symbol is a function of a predetermined set of data symbols [9], [10], [11], [12], [13], [14], [15], [16], [17]. Another recent paper [18] represents code symbols as vertices of a partially connected graph. Each symbol is a function of its neighbors and, if erased, can be recovered from them. Our code also utilizes a graph structure, though only to describe the encoding procedure. There is not necessarily a notion of an individual code symbol being repairable from a designated local subset of the other code symbols.

II. Problem Setup

Consider a bipartite graph $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ with $s = |\mathcal{M}| \leq |\mathcal{V}| = n$. The set $\mathcal{E}$ is the set of edges of the graph, with $(m_i, c_j) \in \mathcal{E}$ if and only if $m_i \in \mathcal{M}$ is connected to $c_j \in \mathcal{V}$. This graph defines a code where the vertices $\mathcal{M}$ correspond to message symbols and the vertices $\mathcal{V}$ correspond to codeword symbols. A bipartite graph with $s = 3$ and $n = 7$ is depicted in figure 1. Thus, if each m_i and c_j are assigned values in the finite field $\mathbb{F}_q$ with q elements, then our messages are the vectors $\mathbf{m} = (m_1, \ldots, m_s) \in \mathbb{F}_q^s$ and our codewords are the vectors $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{F}_q^n$. Each codeword symbol c_j will be a function of the message symbols to which it is connected, as we will now formalize.

Henceforth, $[\mathbf{c}]_{\mathcal{I}}$ is the subvector of $\mathbf{c}$ with elements indexed by $\mathcal{I} \subseteq \{1, ..., n\}$, and $[\mathbf{A}]_{i,j}$ is the $(i,j)^{\text{th}}$ element of a matrix $\mathbf{A}$. Let $\mathcal{N}(c_j)$ denote the neighborhood of $c_j \in \mathcal{V}$, i.e. $\mathcal{N}(c_j) = \{m_i \in \mathcal{M} : (m_i, c_j) \in \mathcal{E}\}$. Similarly, define $\mathcal{N}(m_i) = \{c_j : (m_i, c_j) \in \mathcal{E}\}$. We will also consider neighborhoods of subsets of the vertex sets, i.e. for $\mathcal{V}' \subseteq \mathcal{V}$, $\mathcal{N}(\mathcal{V}') = \cup_{c_j \in \mathcal{V}'} \mathcal{N}(c_j)$. The neighborhood of a subset of $\mathcal{M}$ is defined in a similar manner. Let m_i take values in $\mathbb{F}_q$ and associate with each $c_j \in \mathcal{V}$ a function $f_j : \mathbb{F}_q^s \longrightarrow \mathbb{F}_q$. We restrict each f_j to be a function of $\mathcal{N}(c_j)$ only. Now consider the set $\mathcal{C} = \{(c_1, \ldots, c_n) : c_j = f_j(\mathbf{m}), \mathbf{m} \in \mathbb{F}_q^s\}$. The set $\mathcal{C}$ is an error-correcting code of length n and size at most q^s. We will denote the minimum distance of $\mathcal{C}$ as $d(\mathcal{C})$. If we restrict f_j to be *linear*, then we obtain a linear code with dimension at most s.

The structure of the code's generator matrix can be deduced from the graph G. Let $\mathbf{g}_j \in \mathbb{F}_q^{s \times 1}$ be a column vector such that the i^{th} entry is zero if $m_i \notin \mathcal{N}(c_j)$. Defining $f_j(\mathcal{N}(c_j)) = \mathbf{m}\mathbf{g}_j$ yields a linear function in which

c_j is a function of $\mathcal{N}(c_j)$ only, as required. A concatenation of the vectors $\mathbf{g}_j$ forms the following matrix:

$$\mathbf{G} = \begin{bmatrix} | & & | \\ \mathbf{g}_1 & \cdots & \mathbf{g}_n \\ | & & | \end{bmatrix} \tag{1}$$

where $\mathbf{G} \in \mathbb{F}_q^{s \times n}$ is the generator matrix of the code $\mathcal{C}$.

We associate with the bipartite graph $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ an adjacency matrix $\mathbf{A} \in \{0,1\}^{s \times n}$, where $[\mathbf{A}]_{i,j} = 1$ if and only if $(m_i, c_j) \in \mathcal{E}$. For the example in figure 1, this matrix is equal to

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{2}$$

A *valid* generator matrix $\mathbf{G}$ (in generic form) is built from $\mathbf{A}$ by replacing non-zero entries with indeterminates. The choice of indeterminates (from a suitably-sized finite field $\mathbb{F}_q$) determines the dimension of the code and its minimum distance. For general linear codes, the Singleton bound (on minimum distance) is tight over large alphabets. In the presence of encoding constraints, the Singleton bound can be rather loose. In the next section, we derive an upper bound on the minimum distance of any code (linear or non-linear) associated with a bipartite graph. This bound is reminiscent of the cut-set bounds of Dikaliotis et al. in [1].

A. Subcodes of Reed-Solomon Codes

Throughout this paper, we use the original definition of an $[n,k]_q$ Reed-Solomon code as in [19], the k-dimensional subspace of $\mathbb{F}_q^n$ given by $\mathcal{C}_{\text{RS}} = \{(m(\alpha_1), \ldots, m(\alpha_n)) : \deg(m(x)) < k\}$, where the $m(x)$ are polynomials over $\mathbb{F}_q$ of degree $\deg(m(x))$, and the $\alpha_i \in \mathbb{F}_q$ are distinct (fixed) field elements. Each message vector $\mathbf{m} = (m_0, \ldots, m_{k-1})$ is mapped to a message polynomial $m(x) = \sum_{i=0}^{k-1} m_i x^i$, which is then evaluated at the n elements $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of $\mathbb{F}_q$, known as the defining set of the code. Reed-Solomon codes are MDS codes; their minimum distance attains the Singleton bound, i.e. $d(\mathcal{C}_{\text{RS}}) = n - k + 1$.

We can extract a subcode of a Reed-Solomon code that is valid for the bipartite graph $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ as follows: First, let $\mathbb{F}_q$ be a finite field with cardinality $q \geq n$. Associate to each $c_j \in \mathcal{V}$ a distinct element $\alpha_j \in \mathbb{F}_q$. Consider the i^{th} row of the adjacency matrix $\mathbf{A}$ of G, and let $t_i(x) = \prod_{j:[\mathbf{A}]_{i,j}=0}(x - \alpha_j)$. For example, $t_3(x) = (x - \alpha_1)(x - \alpha_2)$ corresponds to the the third row of $\mathbf{A}$ in (2). Choose k such that $k > \deg(t_i(x))$, $\forall i$. If $\mathbf{t}_i \in \mathbb{F}_q^k$ is the (row) vector of coefficients of $t_i(x)$ and $\mathbf{G}_{\text{RS}}$ is the generator matrix of a Reed-Solomon code with defining set $\{\alpha_1, \ldots, \alpha_n\}$ and dimension k, then $\mathbf{t}_i \mathbf{G}_{\text{RS}} = (t_i(\alpha_1), \ldots, t_i(\alpha_n))$ is a vector that is valid for the i^{th} row of $\mathbf{G}$, i.e. if $[\mathbf{A}]_{i,j} = 0$ then $[\mathbf{t}_i \mathbf{G}_{\text{RS}}]_j = 0$. A horizontal stacking of the vectors $\mathbf{t}_i$ results in a transformation matrix $\mathbf{T}$ that will produce a valid generator matrix $\mathbf{G}$ from $\mathbf{G}_{\text{RS}}$:

$$\mathbf{G} = \mathbf{T}\mathbf{G}_{\text{RS}} = \begin{bmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_s \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ \alpha_1 & \cdots & \alpha_n \\ \vdots & \ddots & \vdots \\ \alpha_1^{(k-1)} & \cdots & \alpha_n^{(k-1)} \end{bmatrix} \tag{3}$$

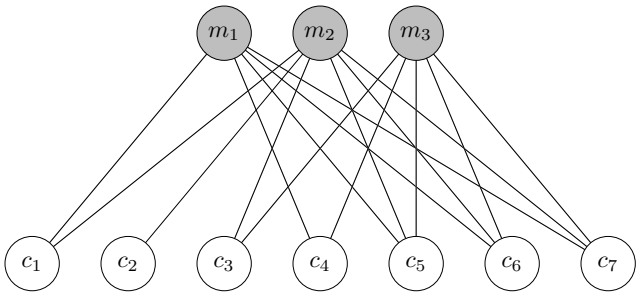

Fig. 1. A bipartite graph representing with 3 message symbols and 7 code symbols

The rank of $\mathbf{G}$ will be equal to the rank of $\mathbf{T}$, and the resulting code $\mathcal{C}$ will have a minimum distance $d(\mathcal{C})$ that is determined by $\mathcal{C}_{\text{RS}}$. Indeed, $d(\mathcal{C}) \geq d(\mathcal{C}_{\text{RS}})$.

III. Minimum Distance

In this section, an upper bound on the minimum distance of a code defined by a bipartite graph $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ is derived. The bound closely resembles the cut-set bounds of [1]. In most cases, this bound is tighter than the Singleton bound for a code of length n and dimension s. For each $\mathcal{M}' \subseteq \mathcal{M}$ define $n_{\mathcal{M}'} := |\mathcal{N}(\mathcal{M}')|$. This is the number of code symbols c_j in $\mathcal{V}$ that are a function of the information symbols $\mathcal{M}'$. The following proposition characterizes the minimum distance of any code defined by G.

Proposition 1. *Fix a field $\mathbb{F}_q$. For any code $\mathcal{C}$ with $|\mathcal{C}| = q^s$ defined by a fixed graph $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$, the minimum distance $d(\mathcal{C})$ obeys*

$$d(\mathcal{C}) \leq n_{\mathcal{M}'} - |\mathcal{M}'| + 1, \quad \forall \mathcal{M}' \subseteq \mathcal{M}. \tag{4}$$

Proof: Working toward a contradiction, suppose $d(\mathcal{C}) > n_{\mathcal{I}} - |\mathcal{I}| + 1$ for some $\mathcal{I} \subseteq \mathcal{M}$. Let $\mathcal{C}'$ be the encoding of all message vectors $\mathbf{m}$ where $[\mathbf{m}]_{\mathcal{I}^{\text{c}}} \in \mathbb{F}_q^{|\mathcal{I}^{\text{c}}|}$ has some arbitrary but fixed value. Note that $[\mathbf{c}]_{\mathcal{N}(\mathcal{I})^{\text{c}}}$ is the same for all $\mathbf{c} \in \mathcal{C}'$, since the symbols $\mathcal{N}(\mathcal{I})^{\text{c}}$ are a function of $\mathcal{I}^{\text{c}}$ only. Since $|\mathcal{I}| > n_{\mathcal{I}} - d(\mathcal{C}) + 1$, then by the pigeonhole principle there exist $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}'$ such that, without loss of generality, the first $n_{\mathcal{I}} - d(\mathcal{C}) + 1$ symbols of $[\mathbf{c}_1]_{\mathcal{N}(\mathcal{I})}$ and $[\mathbf{c}_2]_{\mathcal{N}(\mathcal{I})}$ are identical. Furthermore, $[\mathbf{c}_1]_{\mathcal{N}(\mathcal{I})^{\text{c}}} = [\mathbf{c}_2]_{\mathcal{N}(\mathcal{I})^{\text{c}}}$. Finally, since $\mathcal{N}(\mathcal{I})$ and $\mathcal{N}(\mathcal{I})^{\text{c}}$ partition $\mathcal{V}$, we obtain $d_{\text{H}}(\mathbf{c}_1, \mathbf{c}_2) \leq n - (n_{\mathcal{I}} - d(\mathcal{C}) + 1 + (n - n_{\mathcal{I}})) = d(\mathcal{C}) - 1$, a contradiction. Figure 2 illustrates the relation between $\mathcal{I}$ and the corresponding partition of $\mathcal{V}$. ■

As a direct corollary, we obtain the following upper bound on $d(\mathcal{C})$:

Corollary 1.

$$d(\mathcal{C}) \leq \min_{\mathcal{M}' \subseteq \mathcal{M}} \{n_{\mathcal{M}'} - |\mathcal{M}'|\} + 1 \tag{5}$$

Our next task is to provide constructions of codes that achieve this bound.

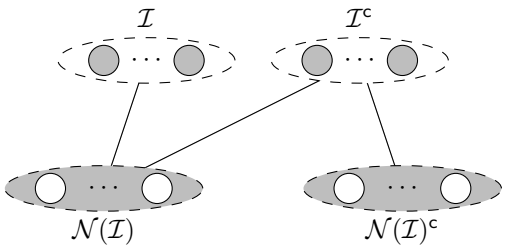

Fig. 2. Partitions of $\mathcal{M}$ and of $\mathcal{V}$ used in the proof of proposition 1. The set $\mathcal{N}(\mathcal{I})$ is a function of both $\mathcal{I}$ and $\mathcal{I}^{\mathsf{c}}$, while the set $\mathcal{N}(\mathcal{I})^{\mathsf{c}}$ is a function of $\mathcal{I}^{\mathsf{c}}$ only.

IV. Systematic Construction

In this section, we provide a code construction that achieves the minimum distance bound stated in corollary 1. We appeal to Hall's Theorem, a well-known result in graph theory that establishes a necessary and sufficient condition for finding a matching in a bipartite graph. Some terminology needed from graph theory is defined in the following subsection.

A. Graph Theory Preliminaries

Let $G = (\mathcal{S}, \mathcal{T}, \mathcal{E})$ be a bipartite graph. A *matching* is a subset $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ such that no two edges in $\tilde{\mathcal{E}}$ share a common vertex. A vertex is said to be *covered* by $\tilde{\mathcal{E}}$ if it is incident to an edge in $\tilde{\mathcal{E}}$. An $\mathcal{S}$*-covering* matching is one by which each vertex in $\mathcal{S}$ is covered. We will abuse terminology and say that an edge $e \in \tilde{\mathcal{E}}$ is *unmatched* if $e \notin \tilde{\mathcal{E}}$. We can now state Hall's Theorem.

Theorem 1. *Let $G = (\mathcal{S}, \mathcal{T}, \mathcal{E})$ be a bipartite graph. There exists an $\mathcal{S}$-covering matching if and only if $|\mathcal{S}'| \leq \mathcal{N}(\mathcal{S}')$ for all $\mathcal{S}' \subseteq \mathcal{S}$.*

For a proof of the theorem, see e.g. [20, p.53].

Set $d_{\mathsf{min}} = \min_{\mathcal{M}' \subseteq \mathcal{M}} \{n_{\mathcal{M}'} - |\mathcal{M}'|\} + 1$. In order to construct a generator matrix $\mathbf{G} \in \mathbb{F}_q^{s \times n}$ for a code $\mathcal{C}$ with minimum distance d_{min}, we will use an $[n, n - d_{\mathsf{min}} + 1]$ Reed-Solomon code with generator matrix $\mathbf{G}_{\mathrm{RS}}$. We will then extract $\mathcal{C}$ as a subcode using an appropriately built transformation matrix $\mathbf{T}$ to form $\mathbf{G} = \mathbf{T}\mathbf{G}_{\mathrm{RS}}$ such that $\mathbf{G}$ is in systematic form, which implies that the dimension of $\mathcal{C}$ is s. Since $\mathcal{C}$ is a subcode of a code with minimum distance d_{min}, we have $d(\mathcal{C}) \geq d_{\mathsf{min}}$. (5) further implies that $d(\mathcal{C}) = d_{\mathsf{min}}$.

Our construction is as follows: consider a graph $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ defining $\mathcal{C}$, and define the set $\mathcal{A} = \{c_j : \mathcal{N}(c_j) = \mathcal{M}\}$, i.e. $\mathcal{A}$ is the set of code symbols that are a function of *every* message symbol. Note that $\mathcal{A} \subseteq \mathcal{N}(\mathcal{M}')$ for every $\mathcal{M}' \subseteq \mathcal{M}$. Therefore, if $a = |\mathcal{A}|$ then the size of the neighborhood of $\mathcal{N}(\mathcal{M}')$ can be expressed as $n_{\mathcal{M}'} = r_{\mathcal{M}'} + a$, where $r_{\mathcal{M}'}$ is the cardinality of the set $\mathcal{R}(\mathcal{M}') = \mathcal{N}(\mathcal{M}') \setminus \mathcal{A}$.

Theorem 2. *Let $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$. Set $d_{\mathsf{min}} = \min_{\mathcal{M}' \subseteq \mathcal{M}} \{n_{\mathcal{M}'} - |\mathcal{M}'|\} + 1$ and $k_{\mathsf{min}} = n - d_{\mathsf{min}} + 1$. A linear code $\mathcal{C}$ with parameters $[n, s, d_{\mathsf{min}}]$ valid for G can be constructed with a systematic-form generator matrix provided that $k_{\mathsf{min}} \geq r_{\mathcal{M}}$.*

Proof: First, we establish a bound on a. Note that since $n = n_{\mathcal{M}} = r_{\mathcal{M}} + a$ and $k_{\min} \geq r_{\mathcal{M}}$, then we have $a \geq d_{\min} - 1$. Fix an arbitrary subset $\mathcal{A}^* \subseteq \mathcal{A}$ of size $a^* = a - (d_{\min} - 1)$, which is guaranteed to exist by virtue of the bound on a, and let $\mathcal{B} = \mathcal{A} \setminus \mathcal{A}^*$. Now, we focus on a particular subgraph of G defined by $G^* = (\mathcal{M}, \mathcal{V}^*, \mathcal{E}^*)$ where $\mathcal{V}^* = \mathcal{V} \setminus \mathcal{B}$, and $\mathcal{E}^* = \{(m_i, c_j) \in \mathcal{E} : c_j \in \mathcal{V}^*\}$ is the edge set corresponding to this subgraph. Since $n_{\mathcal{M}'} = r_{\mathcal{M}'} + a$, then from the definition of $d_{\min}$ we have

$$|\mathcal{M}'| \leq r_{\mathcal{M}'} + a - (d_{\min} - 1), \quad \forall \mathcal{M}' \subseteq \mathcal{M} \tag{6}$$

The neighborhood of every subset $\mathcal{M}'$ when restricted to $\mathcal{V}^*$ is exactly $\mathcal{N}^*(\mathcal{M}') = \mathcal{R}(\mathcal{M}') \cup \mathcal{A}^*$, with cardinality $n^*_{\mathcal{M}'} = r_{\mathcal{M}'} + a^*$. The bounds (6) can now be expressed in a way suitable for the condition of Hall's theorem:

$$|\mathcal{M}'| \leq n^*_{\mathcal{M}'}, \quad \forall \mathcal{M}' \subseteq \mathcal{M} \tag{7}$$

An $\mathcal{M}$-covering matching in G^* can be found by letting $\mathcal{S} = \mathcal{M}$ and $\mathcal{T} = \mathcal{V}^*$ in theorem 1. Let $\tilde{\mathcal{E}} = \{(m_i, c_{j(i)})\}_{i=1}^{s} \subseteq \mathcal{E}^*$ be such a matching, and $\tilde{\mathcal{V}}$ the subset of $\mathcal{V}^*$ that is covered by $\tilde{\mathcal{E}}$. Let $\mathbf{A}_{\tilde{\mathcal{E}}}$ be the adjacency matrix of G when the edge set $\{(m_i, c_j) \in \mathcal{E} : c_j \in \tilde{\mathcal{V}}, j \neq j(i)\}$ is removed. The number of zeros in any row of $\mathbf{A}_{\tilde{\mathcal{E}}}$ is at most $n - d_{\min}$. To see this, note that the edges in $\mathcal{E}$ incident to $\mathcal{B}$ are not removed by the matching, and every $m_i \in \mathcal{M}$ is connected to at least one vertex in $\mathcal{V}^*$. Next, we build a valid $\mathbf{G}$ for G using $\mathbf{A}_{\tilde{\mathcal{E}}}$, utilizing the method described in section II-A. Fix a $[n, n - d_{\min} + 1]$ Reed-Solomon code with generator matrix $\mathbf{G}_{\text{RS}}$ and defining set $\{\alpha_1, \ldots, \alpha_n\}$. The i^{th} transformation polynomial is $t_i(x) = \prod_{j:[\mathbf{A}_{\tilde{\mathcal{E}}}]_{i,j}=0}(x - \alpha_i)$. Since the number of zeros in any row of $\mathbf{A}_{\tilde{\mathcal{E}}}$ is at most $n - d_{\min}$, we have $\deg(t_i(x)) \leq n - d_{\min} = k - 1$ for all i. We use the $t_i(x)$, after normalizing by $t_i(\alpha_{j(i)})$, to construct a transformation matrix $\mathbf{T}$ and then $\mathbf{G} = \mathbf{T}\mathbf{G}_{\text{RS}}$ is valid for G. Note that $\mathbf{G}$ is in systematic form due the fact that the columns of $\mathbf{A}_{\tilde{\mathcal{E}}}$ indexed by $\{j(i)\}_{i=1}^{s}$ form a permutation of the identity matrix of size s. Lastly, $d(\mathcal{C}) = d_{\min}$ since $d(\mathcal{C}) \leq d_{\min}$ by corollary (5), and $d(\mathcal{C}) \geq d_{\min}$ since $\mathcal{C}$ is a subcode of a code with minimum distance $d_{\min}$. ■

V. Minimum Distance for Systematic Linear Codes

In this section, we will restrict our attention to the case where a code valid for G is linear, so that each $c_j \in \mathcal{V}$ is a linear function of the message symbols $m_i \in \mathcal{N}(c_j)$. We seek to answer the following: What is the greatest minimum distance attainable by a *systematic* linear code valid for G?

Any systematic code must correspond to a matching $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ which identifies each message symbol $m_i \in \mathcal{M}$ with a unique codeword symbol $c_{j(i)} \in \mathcal{V}$, where $j(i) \in \{1, \ldots, n\}$. Explicitly, $\tilde{\mathcal{E}}$ consists of s edges of the form $\{(m_i, c_{j(i)})\}$ for $i = 1, \ldots, s$ such that $c_{j(i_1)} \neq c_{j(i_2)}$ for $i_1 \neq i_2$. As before, $\tilde{\mathcal{V}}$ is the subset of vertices in $\mathcal{V}$ which are involved in the matching: $\tilde{\mathcal{V}} = \{c_{j(i)}\}_{i=1}^{s}$. Our code becomes systematic by setting $c_{j(i)} = m_i$ for $i = 1, \ldots, s$, and choosing each remaining codeword symbol $c_j \notin \tilde{\mathcal{V}}$ to be some linear function of its neighboring message symbols $m_i \in \mathcal{N}(c_j)$.

Definition 1. *For $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$, let $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ be an $\mathcal{M}$-covering matching so that $\tilde{\mathcal{E}} = \{(m_i, c_{j(i)})\}_{i=1}^{s}$. Let $\tilde{\mathcal{V}} = \{c_{j(i)}\}_{i=1}^{s}$ be the vertices in $\mathcal{V}$ which are covered by $\tilde{\mathcal{E}}$. Define the matched adjacency matrix $\mathbf{A}_{\tilde{\mathcal{E}}} \in \{0, 1\}^{s \times n}$*

so that $[\mathbf{A}_{\tilde{\mathcal{E}}}]_{i,j} = 1$ *if and only if either* $(m_i, c_j) \in \tilde{\mathcal{E}}$*, or* $c_j \notin \tilde{\mathcal{V}}$ *and* $(m_i, c_j) \in \mathcal{E}$*. In other words,* $\mathbf{A}_{\tilde{\mathcal{E}}}$ *is the adjacency matrix of the bipartite graph formed by starting with* G *and deleting the edges* $\{(m_i, c_j) \in \mathcal{E} \ : \ c_j \in \tilde{\mathcal{V}}$ *and* $j \neq j(i)\}$*.*

Definition 2. *Let* $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ *be a matching for the* $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ *which covers* $\mathcal{M}$*. Let* $z_{\tilde{\mathcal{E}}}$ *be the maximum number of zeros in any row of the corresponding matched adjacency matrix* $\mathbf{A}_{\tilde{\mathcal{E}}}$*, and define* $k_{\tilde{\mathcal{E}}} := z_{\tilde{\mathcal{E}}} + 1$*. Furthermore, define* $k_{\mathsf{sys}} = min_{\tilde{\mathcal{E}}} k_{\tilde{\mathcal{E}}}$ *where* $\tilde{\mathcal{E}}$ *ranges over all matchings for* G *which cover* $\mathcal{M}$*, and* $d_{\mathsf{sys}} = n - k_{\mathsf{sys}} + 1$*.*

Lemma 1. *For a given bipartite graph* $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ *which merits a matching that covers* $\mathcal{M}$*, we have*

$$s \leq k_{\mathsf{min}} \leq k_{\mathsf{sys}} \leq n \tag{8}$$

and

$$d_{\mathsf{sys}} \leq d_{\mathsf{min}}. \tag{9}$$

Proof: Let $\mathbf{A}$ be the adjacency matrix of G.

For any subset $\mathcal{M}' \subseteq \mathcal{M}$ we have $d_{\mathsf{min}} \leq n_{\mathcal{M}'} - |\mathcal{M}'| + 1$, and likewise $k_{\mathsf{min}} = n - d_{\mathsf{min}} + 1 \geq |\mathcal{M}'| + (n - n_{\mathcal{M}'})$. Taking $\mathcal{M}' = \mathcal{M}$ (and noting that in our framework, every $c_j \in \mathcal{V}$ is connected to at least one vertex in $\mathcal{M}$, hence $n_{\mathcal{M}} = n$) we obtain $k_{\mathsf{min}} \geq s$.

Now choose a set $\mathcal{M}'$ for which the above relation holds with equality, that is, $k_{\mathsf{min}} = |\mathcal{M}'| + (n - n_{\mathcal{M}'})$. Since $\mathcal{N}(\mathcal{M}')$ is simply the union of the support sets of the rows of $\mathbf{A}$ corresponding to $\mathcal{M}'$, then each of these rows must have at least $n - n_{\mathcal{M}'} = |\mathcal{N}(\mathcal{M}')^{\mathsf{c}}|$ zeros. Furthermore, any matching $\tilde{\mathcal{E}}$ which covers $\mathcal{M}$ must identify the rows of $\mathcal{M}'$ with columns of $\mathcal{N}(\mathcal{M}')$. Thus, in the matched adjacency matrix $\mathbf{A}_{\tilde{\mathcal{E}}}$, the row corresponding to $j \in \mathcal{M}'$ must have $|\mathcal{M}'| - 1$ zeros in the columns of $\mathcal{N}(\mathcal{M})$ which are matched to $\mathcal{M}' \setminus \{j\}$, in addition to the $n - n_{\mathcal{M}'}$ zeros in the columns corresponding to $\mathcal{N}(\mathcal{M}')$. This gives us $k_{\tilde{\mathcal{E}}} \geq |\mathcal{M}'| + (n - n_{\mathcal{M}'})$ for each matching $\tilde{\mathcal{E}}$, hence $k_{\mathsf{sys}} \geq k_{\mathsf{min}}$. It follows directly that $d_{\mathsf{sys}} \leq d_{\mathsf{min}}$. Finally, it is clear from definition that for any $\mathcal{M}$-covering matching $\tilde{\mathcal{E}}$ we must have that $k_{\tilde{\mathcal{E}}}$ is less than the length of the adjacency matrix $\mathbf{A}$, which is n, hence $k_{\mathsf{sys}} \leq n$. ■

Corollary 1. *Let* $G = (\mathcal{M}, \mathcal{V}, \mathcal{E})$ *be a bipartite graph which merits a systematic linear code. The largest minimum distance obtainable by a systematic linear code is* d_{sys}*.*

Proof: Let $\mathcal{C}$ be a systematic linear code which is valid for G. Then $\mathcal{C}$ must have a codeword containing at least $k_{\mathsf{sys}} - 1$ zeros, i.e. a codeword of Hamming weight at most $n - k_{\mathsf{sys}} + 1 = d_{\mathsf{sys}}$. Since the code is linear, this Hamming weight is an upper bound for its minimum distance, so $d(\mathcal{C}) \leq d_{\mathsf{sys}}$.

It remains to see that there are systematic linear codes which are valid for G and achieve a minimum distance of d_{sys}. Let $\tilde{\mathcal{E}}$ be an $\mathcal{M}$-covering matching for G such that $k_{\tilde{\mathcal{E}}} = k_{\mathsf{sys}}$. Then for any $k \geq k_{\mathsf{sys}}$, we claim that an $[n, k]$ Reed-Solomon code contains a systematic linear subcode that is valid for G. Indeed, choose a set of n distinct elements $\{\alpha_i\}_{i=1}^{n} \subseteq \mathbb{F}_q$ as the defining set of our Reed-Solomon code. Then to form our subcode's generator matrix $\mathbf{G}$, note that (as mentioned before) $\mathbf{G}$ must have zero entries in the same positions as the zero entries of $\mathbf{A}_{\tilde{\mathcal{E}}}$, and indeterminate elements in the remaining positions. There are at most $k_{\mathsf{sys}} - 1$ zeros in any row of $\mathbf{A}_{\tilde{\mathcal{E}}}$ (and at *least*

$s-1$ zeros in each row, since there must be s columns which have nonzero entries in exactly one row). For each row $i \in \{1,\dots,s\}$ of $\mathbf{A}_{\tilde{\mathcal{E}}}$, let $\mathcal{I}_i \subseteq \{1,\dots,n\}$ be the set of column indices j such that $[\mathbf{A}_{\tilde{\mathcal{E}}}]_{i,j} = 0$. Then form the polynomial $t_i(x) = \prod_{j\in\mathcal{I}_i}(x-\alpha_j)$ and normalize by $t_i(\alpha_{j(i)})$, which accordingly has degree at most k_{sys} (and at least $s-1$). We now set the i^{th} row of $\mathbf{G}$ to be $(t_i(\alpha_1),\dots,t_i(\alpha_n))$, and we see that by construction this row has zeros precisely at the indices $j \in \mathcal{I}_i$ as desired.

The rows of $\mathbf{G}$ generate a code with minimum distance at least that of the original Reed-Solomon code, which is $n-k+1$. Furthermore, by setting $k = k_{\mathsf{sys}}$ for our Reed-Solomon code, we see this new code $\mathcal{C}$ has minimum distance at least $n-k_{\mathsf{sys}}+1 = d_{\mathsf{sys}}$. Since by our previous argument, $d(\mathcal{C}) \leq d_{\mathsf{sys}}$, the minimum distance of $\mathcal{C}$ must achieve d_{sys} with equality. ■

VI. Achievability Using MDS Codes

Throughout this paper, we have utilized Reed-Solomon codes to construct systematic linear codes valid for a particular $G = (\mathcal{M},\mathcal{V},\mathcal{E})$ that attain the highest possible distance. It is worth mentioning that this choice is not necessary and in fact, the Reed-Solomon code utilized can be replaced with any linear MDS code with the same parameters.

Lemma 2. *Fix an arbitrary $[n,k]$ linear MDS code $\mathcal{C}$. For any $\mathcal{I} \subseteq [n]$ where $|\mathcal{I}| \leq k-1$, there exists $\mathbf{c} \in \mathcal{C}$ such that $[\mathbf{c}]_{\mathcal{I}} = \mathbf{0}$.*

Proof: Let $\mathbf{G} = [\mathbf{g}_i]_{i=1}^n$ be the generator matrix of $\mathcal{C}$ and let $\mathbf{G}_{\mathcal{I}} = [\mathbf{g}_i]_{i\in\mathcal{I}}$. Since $|\mathcal{I}| \leq k-1$, $\mathbf{G}_{\mathcal{I}}$ has full column rank and so it has a non-trivial left nullspace of dimension $k-|\mathcal{I}|$. If $\mathbf{h}$ is any vector in that nullspace then $\mathbf{c} = \mathbf{h}\mathbf{G}$ is such that $[\mathbf{c}]_{\mathcal{I}} = \mathbf{0}$. ■

Therefore, to produce a valid linear code $\mathcal{C}$ for $G = (\mathcal{M},\mathcal{V},\mathcal{E})$ with $d(\mathcal{C}) = d^*$, where $d^* \leq n_{m_i}$ for all $m_i \in \mathcal{M}$, we fix an arbitrary $[n, n-d^*+1]$ MDS code and then select vectors $\mathbf{h}_1,\dots,\mathbf{h}_s$ such that $\mathbf{h}_i$ is in the left nullspace of $\mathbf{G}_{\mathcal{I}_i}$, where $\mathcal{I}_i = \{j : \mathbf{A}_{i,j} = 0\}$. Note that the specific selection of the $\mathbf{h}_i$ determines the dimension of $\mathcal{C}$. For a systematic construction, in which the dimension of the code is guaranteed to be s, some extra care has to be taken when choosing the $\mathbf{h}_i$. We must choose each $\mathbf{h}_i$ such that its not in the nullspace of $\mathbf{g}_{j(i)}$, which the column corresponding to the systematic coordinate $c_{j(i)}$.

VII. Example

In this section, we construct a systematic linear code that is valid for the graph in figure 1. The bound of corollary 5 asserts that $d(\mathcal{C}) \leq 5$ for any $\mathcal{C}$ valid for G. However, lemma 1 shows that $d(\mathcal{C}_{\mathsf{sys}}) \leq 4$ for any valid systematic linear code $\mathcal{C}_{\mathsf{sys}}$. A matching achieving this bound is given by the edges $\tilde{\mathcal{E}} = \{(m_1,v_1),(m_2,v_2),(m_3,v_3)\}$ and so the edges removed from the graph are $\{(m_2,v_1),(m_2,v_3)\}$. The new adjacency matrix $\mathbf{A}_{\tilde{\mathcal{E}}}$ is given by,

$$\mathbf{A}_{\tilde{\mathcal{E}}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ \mathbf{0} & 1 & 0 & \mathbf{0} & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{10}$$

where boldface zeros refer to those edges removed from G because of the matching $\tilde{\mathcal{E}}$.

A generator matrix which is valid for $\mathbf{A}_{\tilde{\mathcal{E}}}$ can be constructed from that of a $[7, 4]$ Reed-Solomon code over $\mathbb{F}_7$ with defining set $\{0, 1, \alpha, \ldots, \alpha^5\}$ where α is a primitive element in $\mathbb{F}_7$, using the method described in II-A.

The polynomials corresponding to the transformation matrix are given by,

$$t_1(x) = \alpha^5(x-1)(x-\alpha) \tag{11}$$

$$t_2(x) = \alpha^4 x(x-\alpha)(x-\alpha^2) \tag{12}$$

$$t_3(x) = \alpha^3 x(x-1) \tag{13}$$

Finally, the systematic generator matrix for $\mathcal{C}_{\mathsf{sys}}$ is,

$$\mathbf{G}_{\mathsf{sys}} = \begin{bmatrix} 1 & 0 & 0 & \alpha^2 & \alpha^5 & 1 & \alpha^5 \\ 0 & 1 & 0 & 0 & 1 & \alpha^4 & 1 \\ 0 & 0 & 1 & \alpha^5 & \alpha^5 & \alpha^2 & 1 \end{bmatrix} \tag{14}$$

VIII. Conclusion

In this paper, we have studied the problem of analyzing and designing error-correcting codes when the encoding of every coded symbol is restricted to a subset of the message symbols. We obtain an upper bound on the minimum distance of any such code, similar to the cut-set bounds of [1]. By providing an explicit construction, we show that under certain assumptions this bound is achievable. Furthermore, the field size required for the construction scales linearly with the code length. The second bound is on the minimum distance of linear codes with encoding constraints when the generator matrix is required to be in systematic form. We provide a construction that always achieves this bound. Since all of our constructions are built as subcodes of Reed-Solomon codes, they can be decoded efficiently using standard Reed-Solomon decoders. For future work, it remains to show that the first upper bound is achievable in general over small fields.

References

[1] T. K. Dikaliotis, T. Ho, S. Jaggi, S. Vyetrenko, H. Yao, M. Effros, J. Kliewer, and E. Erez, "Multiple-Access Network Information-Flow and Correction Codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 1067–1079, Feb. 2011.

[2] S. H. Dau, W. Song, Z. Dong, and C. Yuen, "Balanced Sparsest generator matrices for MDS codes," in *Inf. Theory Proc. (ISIT), 2013 IEEE Int. Symp.*, 2013, pp. 1889–1893.

[3] S. H. Dau, W. Song, and C. Yuen, "On the existence of MDS codes over small fields with constrained generator matrices," in *Inf. Theory (ISIT), 2014 IEEE Int. Symp.*, Jun. 2014, pp. 1787–1791.

[4] ——, "On Simple Multiple Access Networks," *IEEE J. Sel. Areas Commun.*, vol. 8716, no. 0733, pp. 1–1, 2014.

[5] M. Yan and A. Sprintson, "Weakly Secure Network Coding for Wireless Cooperative Data Exchange," in *Glob. Telecommun. Conf. (GLOBECOM 2011), 2011 IEEE*, Dec. 2011, pp. 1–5.

[6] M. Yan, A. Sprintson, and I. Zelenko, "Weakly Secure Data Exchange with Generalized Reed Solomon Codes," 2014, pp. 1366–1370.

[7] W. Halbawi, T. Ho, H. Yao, and I. Duursma, "Distributed reed-solomon codes for simple multiple access networks," in *2014 IEEE Int. Symp. Inf. Theory*. IEEE, Jun. 2014, pp. 651–655.

[8] W. Halbawi, T. Ho, and I. Duursma, "Distributed gabidulin codes for multiple-source network error correction," in *2014 Int. Symp. Netw. Coding*. IEEE, Jun. 2014, pp. 1–6.

[9] J. Han and L. A. Lastras-Montano, "Reliable Memories with Subline Accesses," in *2007 IEEE Int. Symp. Inf. Theory*. IEEE, Jun. 2007, pp. 2531–2535.

[10] C. Huang, M. Chen, and J. Li, "Pyramid Codes: Flexible Schemes to Trade Space for Access Efficiency in Reliable Data Storage Systems," in *Sixth IEEE Int. Symp. Netw. Comput. Appl. (NCA 2007)*. IEEE, Jul. 2007, pp. 79–86.

[11] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the Locality of Codeword Symbols," *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6925–6934, Nov. 2012.

[12] D. S. Papailiopoulos and A. G. Dimakis, "Locally repairable codes," in *2012 IEEE Int. Symp. Inf. Theory Proc.* IEEE, Jul. 2012, pp. 2771–2775.

[13] I. Tamo, D. S. Papailiopoulos, and A. G. Dimakis, "Optimal locally repairable codes and connections to matroid theory," in *2013 IEEE Int. Symp. Inf. Theory*. IEEE, Jul. 2013, pp. 1814–1818.

[14] N. Prakash, G. M. Kamath, V. Lalitha, and P. V. Kumar, "Optimal linear codes with a local-error-correction property," in *2012 IEEE Int. Symp. Inf. Theory Proc.* IEEE, Jul. 2012, pp. 2776–2780.

[15] G. M. Kamath, N. Prakash, V. Lalitha, and P. V. Kumar, "Codes with local regeneration," in *2013 Inf. Theory Appl. Work.* IEEE, Feb. 2013, pp. 1–5.

[16] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath, "Optimal Locally Repairable and Secure Codes for Distributed Storage Systems," Oct. 2012.

[17] I. Tamo and A. Barg, "A family of optimal locally recoverable codes," *Information Theory, IEEE Transactions on*, vol. 60, no. 8, pp. 4661–4676, Aug 2014.

[18] A. Mazumdar, "Storage Capacity of Repairable Networks," *arXiv:1408.4862*, Aug. 2014.

[19] I. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. Soc. Ind. Appl. Math.*, 1960.

[20] J. H. Van Lint and R. M. Wilson, *A Course in Combinatorics*. Cambridge University Press, 2011.